Construction of an Index by Maximization of the Sum of its Absolute Correlation Coefficients with the Constituent Variables

SK Mishra

North-Eastern Hill University, Shillong (India)

25. May 2007

Construction of an index by maximization of the sum of its absolute correlation coefficients with the constituent variables

SK Mishra
Department of Economics
North-Eastern Hill University
Shillong (India)

I. Introduction: On many occasions we need to construct an index that represents a number of variables (indicators). Cost of living index, general price index, human development index, index of level of development, etc are some of the examples that are constructed by a weighted (linear) aggregation of a host of variables. The general formula of construction of such an index (OECD, 2003) may be given as

$$I_i = \sum_{j=1}^{m} w_j x_{ij} \equiv w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im} \;;\; i = 1, 2, ..., n$$

where w_j is the weight assigned to the j^{th} variable, x_j and remains constant over all observations of $x_j = (x_{1j}, x_{2j}, ..., x_{nj})$.

The weights, $w = (w_1, w_2, ..., w_m)$, are determined by the importance assigned to the variables $x_j; j = 1, 2, ..., m$. The criterion on which importance of a variable (vis-à-vis other variables) is determined may be varied and usually has its own logic (Munda and Nardo, 2005). For example, in constructing a cost of living index importance of a commodity is determined by the proportion of consumption expenditure allocated to that particular commodity and in constructing the human development index variables such as literacy, life expectancy or income are weighted according to the importance assigned to them in accordance with their perceived roles in determining human development status.

In many cases, however, the analyst does not have any preferred means or logic to determine the relative importance of different variables. In such cases, weights are assigned mathematically. One of the methods to determine such mathematical weights is the Principal Components analysis (McCracken, 2000).

In the Principal Components analysis (Kendall & Stuart, 1968, pp. 285-299) weights are determined such that the sum of the squared correlation coefficients of the index with the constituent variables (used to construct the index) is maximized. In other words, weights in $I = \sum_{j}^{m} w_j x_j$ are determined such that $\sum_{j=1}^{m} r^2(I, x_j)$ is maximized. Here $r(I, x_j)$ is the coefficient of correlation between the index I and the variable x_j.

The Principal Components analysis is a very well established statistical method that has excellent mathematical properties. From $x = (x_1, x_2, ..., x_m)$ one may obtain m (or fewer) indices that are orthogonal with each other. These indices together explain cent percent variation in the original variables $x = (x_1, x_2, ..., x_m)$. Moreover, the first Principal Component (often used to make a single index) explains the largest proportion of variation in the variables $x = (x_1, x_2, ..., x_m)$.

II. Some Practical Problems with the Principal Components Analysis: Although the Principal Components analysis has excellent mathematical properties, one may face some difficulties in using it if one desires to construct a single index of the variables that are not very highly correlated among themselves. The method has a tendency to pick up the subset of highly correlated variables to make the first component, assign marginal weights to relatively poorly correlated subset of variables and/or relegate the latter subset to construction of the subsequent principal components. Now if one has to construct a single index, such an index undermines the poorly correlated set of variables. As a result, practically speaking, the index so constructed is the weighted aggregation of only the preferred (highly correlated) set of variables. In this sense, the index so constructed is *elitist* in nature that has a preference to the highly correlated subset over the poorly correlated subset of variables. Further, since there is no dependable method available to obtain a composite index by merging two or more principal components, the deferred set of variables never finds its representation in the further analysis.

III. A Wider View of Constructing an Index: Let us now investigate into the possibilities of maximizing $\left(\sum_{j=1}^{m}\left|r(I,x_j)\right|^{L}\right)^{1/L}$ to obtain weights to construct $I=\sum_{j}^{m}w_jx_j$. This is only a Minkowsky generalization of maximization of $\sum_{j=1}^{m}r^2(I,x_j)$ or (equivalently) $\left(\sum_{j=1}^{m}\left|r(I,x_j)\right|^{2}\right)^{1/2}$. It can be shown that as L becomes larger, the index becomes more and more *egalitarian* with an ever-stronger tendency to assign weights such that all or most of the variables are equally correlated with the index. In so doing, it maximizes the minimal correlation of the index with its constituent variables or in other words it gives us the maximin index. However, for $L=1$ the index is *inclusive* in nature that assigns reasonable (although smaller) weights to the members of less correlated subset of variables, but has no tendency to undermine the less correlated variables and their representation. This property of the index obtained by maximizing $\sum_{j=1}^{m}\left|r(I,x_j)\right|$ or maximizing the minimal correlation is attractive and useful. The objective of this paper is to illustrate this fact.

IV. An Experiment: We have conducted (limited) experiments on constructing indices by maximizing (a) sum of squared correlation, which is the standard Principal Components analysis, (b) maximin correlation, and (c) maximizing the sum of absolute correlations. For sake of identification, we would call them I-2, I-M and I-1 respectively. The experiments have been conducted for (i) highly correlated variables and (ii) poorly correlated variables.

V. The Method of Optimization: The method of constructing indices by the Principal Components is available in many software packages such as STATISTICA or SPSS. However, the method to construct indices by maximin correlation or maximization of the sum of absolute correlation is not available. We have obtained all indices (I-2, I-M and I-1) by solving max $\left(\sum_{j=1}^{m}\left|r(I,x_j)\right|^{L}\right)^{1/L}$ such that $I=\sum_{j}^{m}w_jx_j$ where w_j are the decision variables. It

is an intricate non-linear optimization problem. Any powerful non-linear programming software may possibly be used for optimization (see Kuester and Mize, 1973 for classical methods and FORTRAN programs). However, we have used the Differential Evolution (DE) method of Global Optimization (which is in the broader family of the Genetic algorithms). The optimization may also be done by the Particle Swarm method often used in Artificial Intelligence (see Mishra, 2006). We have found that the Repulsive Particle Swarm (RPS) method performs as effectively as the Differential Evolution method. We have not presented the results of the RPS optimization to avoid duplication of results. The FORTRAN codes of DE or RPS may be obtained from the author on request.

VI. Findings: The results of our experiments are presented in tables 1 through 2-c. It is evident from the correlation matrices associated with tables 1 through 2-c that in case of highly correlated variables [Table-1(i)], all the three methods have a tendency to yield indices that represent all the constituent variables. However, when the variables are poorly correlated [Tables 2-a(i) onwards], the principal component index (I-2) has a tendency to undermine some variables by poorly correlating with them (and thus not representing them, or relegating them to be represented by the subsequent principal components). On the contrary, I-M and I-1 assign reasonable weights to those variables and thus includes them. Nevertheless, it may be noted that I-1 and I-M pay the cost in terms of the explained variance [sum of squared $r(I, x_j)$] in the constituent variables.

VI. Concluding Remarks: In this exercise we have shown that the principal component indices are *elitist* and they have a tendency to undermine the importance of poorly correlated variables. On the other hand, I-1 is more *inclusive*, and has a tendency to represent even the poorly correlated variables. The I-M indices are *egalitarian* in nature.

It would depend on the analyst whether he is interested in *egalitarian*, *inclusive* or *elitist* method of constructing indices when the constituent variables are not very highly correlated among themselves. This paper has opened up the option to choose the method of constructing a desired type of index.

References

- Kendall, MG and Stuart, A (1968): *The Advanced Theory of Statistics*, Charles Griffin & Co. London, vol. 3.
- Kuester, J.L. and Mize, J.H. (1973): *Optimization Techniques with Fortran*, McGraw-Hill Book Co. New York.
- McCracken, K (2000): "Some Comments on the Seifa96 Indexes", Paper presented in the 10th *Biennial Conference of the Australian Population Association*, Melbourne, Australia. www.apa.org.au/upload/2000-7B_McCracken.pdf
- Mishra, SK (2006): "Global Optimization by Differential Evolution and Particle Swarm Methods: Evaluation on Some Benchmark Functions". SSRN http://ssrn.com/abstract=933827
- Munda, G. and Nardo, M (2005) : "Constructing Consistent Composite Indicators: The Issue of Weights", EUR 21834 EN, Institute for the Protection and Security of the citizen, European Commission, Luxembourg.
- OECD (2003) *Composite Indicators of Country Performance: A Critical Assessment*, DST/IND(2003)5, Paris.

Table-1(i): Construction of Indices with Highly Correlated Variables							
x_1	x_2	x_3	x_4	x_5	I-2	I-M	I-1
0.24746	0.62495	0.64798	0.29265	0.31671	0.935329	0.555104	0.931487
0.06005	0.04168	0.04671	0.08230	0.08601	0.138445	0.123705	0.139382
0.21551	0.22392	0.24862	0.43289	0.20651	0.559913	0.453251	0.562815
0.00467	0.00204	0.00207	0.00349	0.00036	0.005571	0.002652	0.005566
0.00492	0.00094	0.00148	0.00339	0.06350	0.035359	0.053324	0.035988
0.00000	0.00000	0.00000	0.00000	0.05235	0.025221	0.042251	0.025738
0.08357	0.05477	0.05515	0.09097	0.04490	0.143321	0.098184	0.143692
0.01201	0.00437	0.00596	0.01215	0.05122	0.039473	0.048957	0.039997
0.37148	0.48721	0.51445	0.86138	0.68763	1.243983	1.134537	1.252297
0.13528	0.13168	0.14674	0.25535	0.21537	0.379531	0.342782	0.382132
0.03036	0.02167	0.02812	0.05547	0.03834	0.074205	0.066094	0.074761
0.00517	0.00247	0.00273	0.00486	0.00093	0.007031	0.003961	0.007038
0.22977	0.23106	0.24024	0.39711	0.32794	0.613936	0.533301	0.617538
0.10075	0.13038	0.13673	0.22851	0.09439	0.289373	0.230178	0.290689
0.24962	0.29629	0.32783	0.57561	0.25105	0.71175	0.583349	0.715577
1.00000	1.00000	1.00000	0.99986	1.00000	2.209297	1.576851	2.212113
0.09500	0.09544	0.11392	0.21605	0.07342	0.246902	0.198637	0.248312
0.00692	0.00531	0.00670	0.01292	0.04406	0.034314	0.043812	0.034788
0.15371	0.23098	0.24742	0.41749	0.12368	0.484885	0.379555	0.487018
0.12861	0.14135	0.16284	0.29275	0.12030	0.353504	0.288644	0.355469
0.48198	0.64806	0.63219	0.99368	0.45382	1.354525	1.054034	1.359591
0.38457	0.58160	0.60495	1.00000	0.33558	1.206889	0.947206	1.212009
0.26555	0.46573	0.46191	0.73598	0.18568	0.873286	0.65606	0.876124
0.00167	0.00766	0.01020	0.02037	0.00000	0.015257	0.012841	0.015369
Coefficient of correlation of x_i with the Index					**SAR**	**SSR**	**Index**
0.974402	0.982644	0.982673	0.93019	0.946579	4.816488	4.641960	**I-2**
0.955934	0.953653	0.955266	0.953653	0.953653	4.772161	4.554708	**I-M**
0.974231	0.982138	0.982199	0.931172	0.946778	4.816518	4.641905	**I-1**
SAR=Sum of absolute correlation coefficients; SSR=Sum of squared correlation coefficients							

Table-1(ii): Correlation among Variables and Indices [Ref. Table-1(i)]								
Variable	X1	X2	X3	X4	X5	I_2	I_M	I_1
X1	1.00	.94	.94	.87	.95	.97	.96	.97
X2	.94	1.00	1.00	.90	.89	.98	.95	.98
X3	.94	1.00	1.00	.90	.89	.98	.96	.98
X4	.87	.90	.90	1.00	.82	.93	.95	.93
X5	.95	.89	.89	.82	1.00	.95	.95	.95
I_2	.97	.98	.98	.93	.95	1.00	.99	1.00
I_M	.96	.95	.96	.95	.95	.99	1.00	.99
I_1	.97	.98	.98	.93	.95	1.00	.99	1.00
Non-unitary correlation coefficients in the red are statistically significant at 5% probability level.								

Table-2-a(i): Construction of Indices with Poorly Correlated Variables							
x_1	x_2	x_3	x_4	x_5	I-2	I-M	I-1
0.08073	1.00000	0.82866	0.98526	0.53466	0.949350	0.88374	0.622279
0.16386	0.33756	0.34375	0.81953	0.34072	0.688276	0.456588	0.577067
0.04837	0.41497	0.20507	0.75169	0.07558	0.498488	0.524258	0.262934
0.00000	0.53007	0.33942	0.62952	0.69993	0.461000	0.231441	0.552044
0.16244	0.00674	0.06701	0.37364	0.40331	0.315061	0.006412	0.468762
0.73579	0.85834	0.08372	0.00382	0.44343	0.311464	0.594906	0.214545
0.12815	0.65719	0.68631	0.14273	0.81714	0.473167	0.239402	0.559323
0.71773	0.63278	0.00000	0.00000	0.78624	0.267037	0.251872	0.465655
0.73432	0.92942	0.99401	0.94753	0.32142	1.396340	1.291884	0.851600
0.34473	0.06267	0.34358	0.75008	0.53330	0.778952	0.260126	0.853248
0.73485	0.41459	0.68318	0.52677	0.50320	1.037516	0.675251	0.888701
0.64057	0.00000	0.35274	0.81192	0.67010	0.978199	0.309314	1.108440
0.70792	0.04819	0.91383	1.00000	0.50611	1.475360	0.688011	1.319778
0.33332	0.22854	0.44989	0.07053	0.60184	0.455263	0.114747	0.586056
0.53096	0.28785	1.00000	0.20606	1.00000	0.968056	0.236289	1.140211
1.00000	0.42811	0.85617	0.84209	0.50033	1.455112	0.962163	1.184820
0.24826	0.91986	0.38434	0.35220	0.52443	0.424190	0.573477	0.291396
0.34633	0.43334	0.45968	0.79915	0.57212	0.828557	0.496313	0.779691
0.02417	0.43494	0.66185	0.21522	0.60198	0.478265	0.201933	0.510606
0.85632	0.67677	0.95255	0.45753	0.00000	1.217357	1.203213	0.617591
0.76260	0.64437	0.45344	0.74621	0.40213	0.991050	0.872141	0.738663
0.42520	0.32623	0.97347	0.63480	0.60728	1.129397	0.555505	1.017350
0.84625	0.13717	0.79273	0.35542	0.84484	1.104543	0.365483	1.224984
0.07017	0.68011	0.05396	0.14215	0.25842	0.045625	0.332416	-0.043296
Coefficient of correlation of x_i with the Index					**SAR**	**SSR**	**Index**
0.643674	-0.15995	0.821702	0.671311	-0.02572	2.322359	1.566414	**I-2**
0.52381	0.52381	0.52381	0.52381	-0.52381	2.619051	1.371886	**I-M**
0.573709	-0.5201	0.690693	0.474001	0.421365	2.679864	1.478924	**I-1**
SAR=Sum of absolute correlation coefficients; SSR=Sum of squared correlation coefficients							

Table-2-a(ii): Correlation among Variables and Indices [Ref. Table-2-a(i)]								
Variable	X1	X2	X3	X4	X5	I_2	I_M	I_1
X1	1.00	−.06	.36	.09	.03	.64	.52	.57
X2	−.06	1.00	.02	−.12	−.24	−.16	.52	−.52
X3	.36	.02	1.00	.36	.15	.82	.52	.69
X4	.09	−.12	.36	1.00	−.30	.67	.52	.47
X5	.03	−.24	.15	−.30	1.00	−.03	−.52	.42
I_2	.64	−.16	.82	.67	−.03	1.00	.67	.85
I_M	.52	.52	.52	.52	−.52	.67	1.00	.20
I_1	.57	−.52	.69	.47	.42	.85	.20	1.00
Non-unitary correlation coefficients in the red are statistically significant at 5% probability level.								

Table-2-b(i): Construction of Indices with Poorly Correlated Variables							
x_1	x_2	x_3	x_4	x_5	I-2	I-M	I-1
0.03797	0.91295	0.70434	1.00000	0.49744	-0.377456	-0.313175	0.074617
0.19598	0.37965	0.36607	0.88580	0.36391	-0.456142	-0.060567	0.241395
0.00942	0.40430	0.14686	0.69451	0.05313	-0.215007	-0.038796	0.151148
0.01803	0.60370	0.37592	0.70113	0.76214	-0.115051	-0.534215	-0.213207
0.21162	0.03058	0.07543	0.41734	0.43553	-0.215097	-0.136851	0.037392
0.89672	0.96411	0.09436	0.00825	0.48048	0.095979	-0.481693	-0.272158
0.14601	0.72560	0.74141	0.13223	0.87076	-0.054122	-0.601752	-0.344503
0.87142	0.71549	0.00000	0.00000	0.83681	0.134624	-0.650932	-0.415622
0.77821	0.88109	0.93060	0.76990	0.16688	-0.810971	0.205915	0.568509
0.38798	0.04869	0.33312	0.75149	0.52584	-0.569279	0.012004	0.301019
0.88607	0.47067	0.74681	0.56995	0.54667	-0.755081	0.070259	0.416236
0.78158	0.02269	0.39040	0.90300	0.73099	-0.806088	0.060064	0.443793
0.79128	0.00000	0.93191	0.98178	0.46449	-1.242245	0.479352	0.874433
0.38523	0.23089	0.45373	0.00662	0.63262	-0.198419	-0.253198	-0.087165
0.57407	0.24109	0.99833	0.04206	1.00000	-0.538666	-0.237488	0.038475
0.89741	0.16238	0.61940	0.60711	0.26077	-0.900516	0.381035	0.667035
0.28562	1.00000	0.38885	0.32317	0.55839	0.067499	-0.572907	-0.313645
0.43022	0.49662	0.50713	0.88616	0.61656	-0.533080	-0.164008	0.209251
0.00000	0.42384	0.65143	0.10456	0.62415	-0.114505	-0.369824	-0.190353
1.00000	0.71845	1.00000	0.41461	0.00000	-0.895480	0.427292	0.702464
0.77719	0.51560	0.29537	0.50282	0.31948	-0.435235	-0.014589	0.250614
0.40556	0.18892	0.87828	0.37714	0.56731	-0.664957	0.053448	0.313697
0.94495	0.02000	0.72563	0.15465	0.85811	-0.712214	0.005136	0.282867
0.10263	0.76627	0.05817	0.15495	0.30375	0.260800	-0.499810	-0.362753
Coefficient of correlation of x_i with the Index					**SAR**	**SSR**	**Index**
-0.53918	0.529591	-0.6903	-0.55975	0.203518	2.522336	1.402432	**I-2**
0.513547	-0.51355	0.513547	0.513547	-0.51355	2.567734	1.318651	**I-M**
0.522562	-0.47256	0.563001	0.613438	-0.4413	2.612865	1.384409	**I-1**
SAR=Sum of absolute correlation coefficients; SSR=Sum of squared correlation coefficients							

Table-2-b(ii): Correlation among Variables and Indices [Ref. Table-2-b(i)]

Variable	X1	X2	X3	X4	X5	I_2	I_M	I_1
X1	1.00	−.13	.26	−.08	−.05	−.54	.51	.52
X2	−.13	1.00	−.12	−.17	−.19	.53	−.51	−.47
X3	.26	−.12	1.00	.15	.07	−.69	.51	.56
X4	−.08	−.17	.15	1.00	−.36	−.56	.51	.61
X5	−.05	−.19	.07	−.36	1.00	.20	−.51	−.44
I_2	−.54	.53	−.69	−.56	.20	1.00	−.92	−.97
I_M	.51	−.51	.51	.51	−.51	−.92	1.00	.98
I_1	.52	−.47	.56	.61	−.44	−.97	.98	1.00

Non-unitary correlation coefficients in the red are statistically significant at 5% probability level.

Table-2-c(i): Construction of Indices with Poorly Correlated Variables							
x_1	x_2	x_3	x_4	x_5	I-2	I-M	I-1
0.02807	0.85761	0.64530	1.00000	0.48439	0.062247	-0.445235	-0.218127
0.20525	0.39807	0.36793	0.90801	0.37204	0.179710	-0.282354	-0.094778
0.00294	0.40256	0.12246	0.67552	0.04525	-0.170174	-0.473390	-0.367750
0.03355	0.62184	0.38290	0.73010	0.78395	0.160078	-0.198048	0.033021
0.22661	0.06049	0.07724	0.44322	0.44683	0.254188	0.023270	0.155497
0.91043	0.97513	0.09666	0.03003	0.49347	0.216564	0.035034	0.240758
0.15296	0.73557	0.74887	0.14521	0.88956	0.476663	0.300789	0.492089
0.88394	0.73110	0.00000	0.02033	0.85454	0.399676	0.243724	0.484344
0.75341	0.84125	0.89033	0.70394	0.11269	0.474010	-0.063627	0.122614
0.38888	0.06413	0.32328	0.75292	0.52323	0.480645	0.062668	0.246423
0.89664	0.48940	0.75736	0.59169	0.56191	0.816448	0.325160	0.553455
0.79505	0.05259	0.39756	0.93403	0.75234	0.803562	0.233978	0.497595
0.78126	0.00572	0.92212	0.96992	0.44989	1.018264	0.370059	0.578070
0.38975	0.24277	0.44711	0.00289	0.64341	0.581460	0.464225	0.593489
0.56254	0.23487	0.97997	0.00000	1.00000	1.122808	0.916484	1.097882
0.80662	0.08186	0.51977	0.52392	0.17676	0.694153	0.275357	0.410588
0.29101	1.00000	0.38364	0.32430	0.57030	0.063482	-0.172087	0.028138
0.44443	0.51658	0.51589	0.91611	0.63214	0.425749	-0.096580	0.151960
0.00000	0.42095	0.63596	0.08082	0.63192	0.397961	0.289022	0.405462
1.00000	0.71923	1.00000	0.40818	0.00000	0.710425	0.242741	0.387206
0.74085	0.46585	0.23095	0.41917	0.29050	0.350834	0.020000	0.188367
0.38037	0.15298	0.82704	0.29142	0.55329	0.787776	0.504413	0.637011
0.93079	0.00000	0.68763	0.09593	0.86276	1.190451	0.917797	1.106043
0.11825	0.78049	0.05871	0.17606	0.31965	-0.188370	-0.293398	-0.166479
Coefficient of correlation of x_i with the Index					**SAR**	**SSR**	**Index**
0.639736	-0.58975	0.679074	-0.11904	0.364299	2.391898	1.365094	**I-2**
0.515312	-0.51531	0.515312	-0.51531	0.515312	2.57656	1.327732	**I-M**
0.551179	-0.48884	0.519286	-0.46451	0.572765	2.596579	1.356249	**I-1**
SAR=Sum of absolute correlation coefficients; SSR=Sum of squared correlation coefficients							

Table-2-c(ii): Correlation among Variables and Indices [Ref. Table-2-c(i)]								
Variable	X1	X2	X3	X4	X5	I_2	I_M	I_1
X1	1.00	-.13	.24	-.11	-.05	.64	.52	.55
X2	-.13	1.00	-.14	-.15	-.14	-.59	-.52	-.49
X3	.24	-.14	1.00	.09	.06	.68	.52	.52
X4	-.11	-.15	.09	1.00	-.33	-.12	-.52	-.46
X5	-.05	-.14	.06	-.33	1.00	.36	.52	.57
I_2	.64	-.59	.68	-.12	.36	1.00	.90	.92
I_M	.52	-.52	.52	-.52	.52	.90	1.00	.99
I_1	.55	-.49	.52	-.46	.57	.92	.99	1.00
Non-unitary correlation coefficients in the red are statistically significant at 5% probability level.								